

Submitted to the *Bernoulli*
arXiv: 0909.2139

On the Viterbi process with continuous state space

CHIGANSKY PAVEL^{1,*} RITOV YAACOV^{1,**}

¹*Department of Statistics, The Hebrew University, Mount Scopus, Jerusalem 91905, Israel*
E-mail: ^{*}pchiga@mscc.huji.ac.il; ^{**}yaacov.ritov@gmail.com

This paper deals with convergence of the maximum a posterior probability path estimator in hidden Markov models. We show that when the state space of the hidden process is continuous, the optimal path may stabilize in a way, essentially different from the previously considered finite state setting.

Keywords: Hidden Markov Models, the Viterbi algorithm, MAP path estimator.

1. Introduction

Consider a standard hidden Markov model (X, Y) , where $X = (X_n)_{n \in \mathbb{Z}_+}$ and $Y = (Y_n)_{n \in \mathbb{Z}_+}$ are the *hidden* state and the *observation* processes respectively. The state process X is Markov with values in a subset $\mathcal{S} \subseteq \mathbb{R}$, transition probability Q and initial distribution \mathcal{M} : for all measurable subsets $A \subseteq \mathcal{S}$,

$$\begin{aligned}\mathbb{P}(X_1 \in A) &= \mathcal{M}(A) \\ \mathbb{P}(X_n \in A | X_{n-1}) &= Q(X_{n-1}, A), \quad \mathbb{P} - \text{a.s.} \quad n > 1.\end{aligned}$$

We shall consider either countable \mathcal{S} , in which case $q(u, v) := Q(u, \{v\})$ and $\mu(u) := \mathcal{M}(\{u\})$, or $\mathcal{S} = \mathbb{R}$, assuming that $Q(u, dv)$ and $\mathcal{M}(du)$ have densities $q(u, v)$ and $\mu(u)$ with respect to the Lebesgue measure. The precise meaning of $q(u, v)$ and $\mu(u)$ should be obvious from the context.

The observed process Y forms a sequence of conditionally independent random variables, given $X_{1:\infty} = (X_1, X_2, \dots)$, with the *observation* density p :

$$\mathbb{P}(Y_n \in B | X_{1:\infty}) = \int_B p(X_n, y) dy, \quad \mathbb{P} - \text{a.s.},$$

for any Borel $B \subseteq \mathbb{R}$.

The path estimation problem is to reconstruct the trajectory of the hidden process ¹ $X_{1:n} = (X_1, \dots, X_n)$, given the realization of $Y_{1:n} = (Y_1, \dots, Y_n)$ for a fixed horizon $n \geq 1$.

¹hereafter, for $x \in \mathbb{R}^n$, x_m stands for the m -th entry of x and $x_{k:m}$, $k \leq m$ denotes the vector $x = (x_k, \dots, x_m)$; $|x_{1:n}| = \max_i |x_i|$ and $\|x_{1:n}\| = \sqrt{\sum_{i=1}^n x_i^2}$.

If \mathcal{S} is a discrete set, a natural estimator is the maximizer of the a posteriori probability (MAP estimator):

$$\hat{X}_{1:n}^n := \operatorname{argmax}_{x_{1:n} \in \mathcal{S}^n} \mathbb{P}(X_{1:n} = x_{1:n} | Y_{1:n}),$$

where the optimal path is chosen according to the lexicographical order on \mathcal{S}^n , induced by an order on \mathcal{S} , whenever the maximum is not unique. The obtained path minimizes the probability of error among all estimators depending on $Y_{1:n}$, that is

$$\mathbb{P}(\hat{X}_{1:n}^n \neq X_{1:n}) \leq \mathbb{P}(\xi_{1:n} \neq X_{1:n}), \quad \text{for all } \sigma\{Y_1, \dots, Y_n\}\text{-measurable } \xi_{1:n}.$$

By the Bayes formula,

$$\mathbb{P}(X_{1:n} = x_{1:n} | Y_{1:n}) = \frac{L_n(x_{1:n}, Y_{1:n})}{\sum_{u_{1:n} \in \mathcal{S}^n} L_n(u_{1:n}, Y_{1:n})}$$

where L_n is the “posterior” likelihood:

$$L_n(x_{1:n}; y_{1:n}) = \mu(x_1) p(x_1, y_1) \prod_{m=2}^n q(x_{m-1}, x_m) p(x_m, y_m), \quad x_{1:n} \in \mathcal{S}^n, \quad (1.1)$$

and hence

$$\hat{X}_{1:n}^n = \operatorname{argmax}_{x_{1:n} \in \mathcal{S}^n} L_n(x_{1:n}, Y_{1:n}).$$

Due to the product structure of L_n , the search for the maximizing path can be carried out efficiently by a dynamic programming procedure, called the Viterbi algorithm after A. Viterbi, who introduced it in the context of error correction codes.

When the next observation Y_{n+1} is added, the optimal path may change entirely, i.e., for any $m = 1, \dots, n$, $\hat{X}_{1:m}^{n+1}$ is in general different from $\hat{X}_{1:m}^n$. In practical terms, the latter means that² $\#\mathcal{S}$ optimal pathes candidates of length n are to be kept in memory at each time n . This motivates the question of whether the optimal path stabilizes as the number of observations grows to infinity, or more precisely, whether the limit

$$\hat{X}_{1:m} = \lim_{n \rightarrow \infty} \hat{X}_{1:m}^n \quad (1.2)$$

exists \mathbb{P} -a.s. for each fixed $m \geq 1$. If such a limit exists, it defines a random process with pathes in \mathcal{S}^∞ , coined in [Lember and Koloydenko \(2010\)](#) the *Viterbi process*.

An affirmative answer to this question was given in [Caliebe and Rösler \(2002\)](#) (see also [Kogan \(1996\)](#)) under a sufficient condition (see (2.1) below), which also ensures that the limit sequence $\hat{X} = (\hat{X}_m)_{m \geq 1}$ is a regenerative process. More precisely, a sequence of stopping times can be constructed (see [Caliebe \(2006\)](#)), splitting the process \hat{X} into cycles that are i.i.d. and independent of the initial delay. In particular, by the regenerative property, \hat{X} satisfies the classical limit laws, such as LLN and CLT.

² $\#A$ stands for cardinality of a set A

In fact, the existence of such renewal times under the condition (2.1) can be deduced by a simple argument (replicated for completeness in Section 2). A more delicate construction in Lember and Koloydenko (2008, 2010) verifies (1.2) under conditions, weaker than (2.1).

In this paper, we revisit the question of existence of the limit (1.2) for the hidden Markov models (HMMs) with continuous state space, i.e. when $\mathcal{S} = \mathbb{R}$ and for each $u \in \mathbb{R}$, the transition kernel $Q(u, dv)$ and the initial distribution $\mathcal{M}(dv)$ have densities $q(u, v)$ and $\mu(v)$ with respect to the Lebesgue measure. By the Bayes formula, the conditional law of the vector $X_{1:n}$, given $Y_{1:n}$ has the density ψ_n with respect to the Lebesgue measure on \mathbb{R}^n :

$$\psi_n(x_{1:n}) := \frac{L_n(x_{1:n}; Y_{1:n})}{\int_{\mathbb{R}^n} L_n(u_{1:n}; Y_{1:n}) du_1 \dots du_n},$$

with L_n defined as in (1.1). The MAP path estimator is

$$\hat{X}_{1:n}^n := \operatorname{argmax}_{x_{1:n} \in \mathbb{R}^n} \psi_n(x_{1:n}) = \operatorname{argmax}_{x_{1:n} \in \mathbb{R}^n} L_n(x_{1:n}; Y_{1:n}),$$

where as in (1.2), the maximum is chosen according to the lexicographical order on \mathbb{R}^n (induced e.g. by $<$ on \mathbb{R}) in case of ambiguity.

Note that for any $\sigma\{Y_1, \dots, Y_n\}$ -measurable random vector $\xi_{1:n}$ and $\varepsilon > 0$

$$\mathbb{P}(|X_{1:n} - \xi_{1:n}| \leq \varepsilon) = \mathbb{E}\mathbb{P}(|X_{1:n} - \xi_{1:n}| \leq \varepsilon | Y_{1:n}) = \mathbb{E} \int_{[-\varepsilon, \varepsilon]^n} \psi_n(x_{1:n} + \xi_{1:n}) dx_1 \dots dx_n$$

and hence the estimator $\hat{X}_{1:n}^n$ is optimal in the sense:

$$\lim_{\varepsilon \rightarrow 0} \varepsilon^{-n} \mathbb{P}(|X_{1:n} - \xi_{1:n}| \leq \varepsilon) = \mathbb{E} \psi_n(\xi_{1:n}) \leq \mathbb{E} \max_{x_{1:n} \in \mathbb{R}^n} \psi_n(x_{1:n}) = \lim_{\varepsilon \rightarrow 0} \varepsilon^{-n} \mathbb{P}(|X_{1:n} - \hat{X}_{1:n}^n| \leq \varepsilon),$$

whenever interchanging the expectation and the limit is possible. Roughly this means that $\hat{X}_{1:n}^n$ yields the best “small” credible intervals among all other path estimates³.

As in the state estimation problems such as filtering, the exact calculation of $\hat{X}_{1:n}^n$ is impossible beyond a number of models with a special structure, most notably, Kalman’s linear Gaussian setting. A number of efficient numerical techniques, such as particle filters, have been developed (see e.g. Cappé, Moulines and Rydén (2005)) to approximate the conditional law of the hidden state process. In this paper we are concerned with the convergence properties of the MAP paths, leaving out the computational issues for further investigation.

In Section 2 we explore through a number of examples, various patterns of convergence encountered in (1.2), when the hidden state space is continuous. We also give an example of HMM, for which the MAP path does not converge as the estimation time horizon increases. In Section 3 we prove a more general result, deducing the existence of the

³In fact, this optimality interpretation turns out to be meaningful even in the infinite dimensional function space, see Zeitouni and Dembo (1987, 1988).

limit (1.2) from certain strong log-concavity of the transition and observation densities. Appendix A contains a lemma which is used in the proof of the main result and might be of interest on its own. Finally, a short discussion of the results appears in Section 4.

2. Examples

Let us briefly recall the essential elements of the proof in the finite setting $\mathcal{S} = \{1, \dots, d\}$. For simplicity consider an irreducible finite (and thus recurrent) chain X and define

$$D_i = \{y \in \mathbb{R} : q(x_1, i)p(i, y)q(i, x_3) > q(x_1, x_2)p(x_2, y)q(x_2, x_3), \forall x_2 \neq i, x_1, x_3 \in \mathcal{S}\}.$$

Suppose that for a pair of states j_0 and i_0 ,

$$\int_{D_{i_0}} p(j_0, y)dy > 0. \quad (2.1)$$

Recall the definition of L_n in (1.1) and notice that on the event $A_m = \{X_m = j_0, Y_m \in D_{i_0}\}$ with a fixed $m > 1$ and all $n > m$

$$\begin{aligned} L_n(x_{1:n}, Y_{1:n}) &= L_{m-1}(x_{1:m-1}, Y_{1:m-1})q(x_{m-1}, x_m)p(x_m, Y_m)q(x_m, x_{m+1})L_{m+1,n}(x_{(m+1):n}, Y_{(m+1):n}) \\ &\leq L_{m-1}(x_{1:m-1}, Y_{1:m-1})q(x_{m-1}, i_0)p(i_0, Y_m)q(i_0, x_{m+1})L_{m+1,n}(x_{(m+1):n}, Y_{(m+1):n}), \end{aligned}$$

for an appropriate function $L_{m+1,n}$ and where the equality is attained only at a path $x_{1:m}$ with $x_m = i_0$. Hence the m -th entry of the optimal path must equal i_0 for any $n \geq m$, i.e. $\hat{X}_m^n = i_0$. But then, given \hat{X}_m^n , the first m entries of the optimal path depend only on the values of Y_1, \dots, Y_m and are not affected by $Y_k, k > m$. Hence the limit (1.2) exists on the event A_m . Since the chain (X, Y) is recurrent, for any fixed m , one of the events A_{m+1}, A_{m+2}, \dots occurs \mathbb{P} -a.s. and thus (1.2) holds \mathbb{P} -a.s.

Following the same basic idea, let $\tau(k)$, $k \geq 0$ be the times at which the chain (X, Y) revisits the set $\{j_0\} \times D_{i_0}$:

$$\begin{aligned} \tau(0) &= 1 \\ \tau(k) &= \inf\{n > \tau(k-1) : X_n = j_0, Y_n \in D_{i_0}\}, \quad k \geq 1 \end{aligned}$$

By construction, for any k , on the event $\{\tau(k) \leq n\}$,

$$L(x_{1:n}; Y_{1:n}) \leq L(x_{1:n}^\tau; Y_{1:n}), \quad \forall x_{1:n} \in \mathcal{S}^n,$$

where $x_{1:n}^\tau$ is the vector which coincides with $x_{1:n}$ at all but the indices $\tau(1), \dots, \tau(k)$, where its entries equal i_0 .

The upper bound is attained if $L(x_{1:n}; Y_{1:n})$ is maximized over $x_{1:n}$, constrained to $x_{\tau(1)} = \dots = x_{\tau(k)} = i_0$. Since each $x_{\tau(\ell)}$, $\ell = 1, \dots, k$ appears in the product $L(x_{1:n}; Y_{1:n})$ at three adjacent terms, the optimal choice of each segment $x_{\tau(\ell-1)+1:\tau(\ell)-1}$, $\ell = 1, \dots, k$

is determined only by the values of $Y_{\tau(\ell-1)+1}, \dots, Y_{\tau(\ell)-1}$. Hence, in particular, the limit $\lim_{n \rightarrow \infty} \hat{X}_{1:n}^n$ exists on any of the events $\{\tau(k-1) < m \leq \tau(k) < \infty\}$, $k \geq 1$. By recurrence of j_0 and the condition (2.1), $\mathbb{P}(\tau(k) < \infty) = 1$ and $\lim_{k \rightarrow \infty} \tau(k) = \infty$, \mathbb{P} -a.s., which verifies the existence of the limit (1.2).

The stopping times $\tau(k)$, $k \geq 1$ form a renewal process, with respect to which both (X, Y) and $\hat{X} = (\hat{X}_m)_{m \geq 1}$ are regenerative (see Caliebe (2006) for more details). As pointed out in Lember and Koloydenko (2008) the condition (2.1) can be quite restrictive, especially when the transition matrix is sparse. The convergence in (1.2) and the regenerative property are verified in Lember and Koloydenko (2008) under less conservative conditions, using a more sophisticated construction of the renewal times.

In summary, both Caliebe and Rösler (2002) and Lember and Koloydenko (2008) deduce the existence of the limit in (1.2) from the explicit construction of stopping times, based on the discreteness of the hidden process state space. The following example shows that this still may be possible in HMMs with continuous state space.

Example 2.1. Consider a linear HMM with Laplacian state and Gaussian observation noises:

$$\mu(u) = \frac{1}{4}e^{-|u|/2}, \quad q(u, v) = \frac{1}{4}e^{-|u-v|/2}, \quad p(x, y) = \frac{1}{\sqrt{2\pi}}e^{-(x-y)^2/2}.$$

In this case the MAP path is given by

$$\hat{X}_{1:n}^n = \operatorname{argmin}_{x_{1:n} \in \mathbb{R}^n} \left(|x_1| + (x_1 - Y_1)^2 + \sum_{m=2}^n |x_{m-1} - x_m| + (x_m - Y_m)^2 \right).$$

Consider the function $x \mapsto f(x) := |a - x| + (x - y)^2 + |x - b|$ for fixed $a, b, y \in \mathbb{R}$. Suppose w.l.o.g. $a \leq b$ and note that f , being strictly convex, is minimized at a unique point $x^* = \operatorname{argmin}_{x \in \mathbb{R}} f(x)$. If $y \in [a, b]$, then clearly, $x^* \in [a, b]$ and since on this interval $f(x) = -a + (y - x)^2 + b$, we have $x^* = y$. Consider the case $y \leq a$ and suppose $x^* < a$. For $x < a$, $f(x) = a - x + (y - x)^2 + b - x$ and hence $x^* = y + 1$. By strict convexity, this implies that $x^* = y + 1$, if $y < a - 1$ and that $x^* \geq a$, otherwise. Clearly, $x^* \leq b$, i.e. $x^* \in [a, b]$, which in turn implies that $x^* = a$ for $y \in [a - 1, a)$. Similar calculations reveal that $x^* = y - 1$, if $y > b + 1$ and $x^* = b$ if $y \in (b, b + 1]$.

To summarize, $x^* \in [y - 1, y + 1]$ for any $a, b, y \in \mathbb{R}$ and $x^* = y$, whenever $a \leq y \leq b$. In particular, $\hat{X}_{m-1}^n \in [Y_{m-1} - 1, Y_{m-1} + 1]$ and $\hat{X}_{m+1}^n \in [Y_{m+1} - 1, Y_{m+1} + 1]$ for any $n \geq m + 1$. Hence on the event

$$A_m := \{Y_{m-1} + 1 \leq Y_m \leq Y_{m+1} - 1\},$$

$Y_m \in [\hat{X}_{m-1}^n, \hat{X}_{m+1}^n]$ and consequently $\hat{X}_m^n = Y_m$. This in turn implies $\hat{X}_{1:m}^n = \hat{X}_{1:m}^{m+1}$ for all $n \geq m + 1$ and the existence of the limit (1.2) on any of A_k , $k \geq m + 1$. Clearly A_k 's occur infinitely often and hence, as in the discrete case, $\hat{X}_{1:m}^n$ ceases to change starting from some random but \mathbb{P} -a.s. finite time n . In particular, (1.2) holds \mathbb{P} -a.s. \blacksquare

However, splitting the optimal trajectory into unrelated segments is not the only way to get the convergence in (1.2): the following example shows that the limit may exist without ever being actually attained.

Example 2.2. Consider the linear Gaussian HMM with

$$\mu(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad q(u, v) = \frac{1}{\sqrt{2\pi}} e^{-(u-v)^2/2}, \quad p(x, y) = \frac{1}{\sqrt{2\pi}} e^{-(x-y)^2/2}.$$

In this case the conditional law of $X_{1:n}$, given $Y_{1:n}$ is Gaussian and hence

$$\hat{X}_{1:n}^n = \mathbb{E}(X_{1:n}|Y_{1:n}).$$

For any fixed $m \geq 1$, the process $\hat{X}_{1:m}^n = \mathbb{E}(X_{1:m}|Y_{1:n})$, $n \geq m$ is a uniformly integrable vector valued martingale and hence the limit (1.2) exists by the martingale convergence. In fact, the Kalman linear filtering theory (see e.g. Kwakernaak and Sivan (1972)) tells that in this case (of controllable and observable dynamics) the stronger \mathbb{P} -a.s. exponential convergence holds (see also Remark 3.2 below).

Moreover, $\mathbb{E}(X_{1:m}|Y_{1:n})$ is a deterministic linear map of $Y_{1:n}$ and a calculation reveals that it actually depends on each one of the components in $Y_{1:n}$. Since $Y_{1:n}$ is a non-degenerate Gaussian vector,

$$\mathbb{P}(\hat{X}_j^n = \hat{X}_j^{n'}, \text{ for some } j \leq m) = 0$$

for any $n' > n \geq m$. ■

Finally, the next example demonstrates that a finite limit in (1.2) may not exist, even when the hidden state chain is positive recurrent and has countably many states. In fact, it also shows that the optimal MAP path may not be an adequate estimate: in this case, a trajectory of a positive recurrent chain V is estimated as a constant trajectory, diverging to infinity, as $n \rightarrow \infty$.

Example 2.3. Consider the HMM with the hidden state process $X_n = (U_n, V_n)$, consisting of independent components U and V . The process $U = (U_n)_{n \geq 1}$ is a sequence of i.i.d. random variables uniformly distributed over $[0, 1]$.

$V = (V_n)_{n \geq 1}$ is a random walk on positive integers with reflecting boundary at $\{1\}$ and the transition probabilities $P(1, 1) = 1 - \varepsilon$, $P(1, 2) = \varepsilon$ and for $i \geq 2$,

$$P(i, j) = \begin{cases} \varepsilon \frac{\left(\frac{i}{i+1}\right)^2}{1 + \left(\frac{i}{i+1}\right)^2} & j = i + 1 \\ \varepsilon \frac{1}{1 + \left(\frac{i}{i+1}\right)^2} & j = i - 1 \\ 1 - \varepsilon & j = i \end{cases} \quad (2.2)$$

where $\varepsilon > 0$ is a small fixed constant (in fact, we shall choose $\varepsilon < e^{-2}/(1+e^{-2}) = 0.119\dots$ later on). V is a positive recurrent Markov chain with the unique invariant distribution

$$\pi(j) = \begin{cases} \frac{1}{5}C \left(1 + \left(\frac{1}{2}\right)^2\right) & j = 1 \\ C \frac{1}{j^2} \left(1 + \left(\frac{j}{j+1}\right)^2\right) & j > 1 \end{cases} \quad (2.3)$$

where C is the normalization constant, independent of ε . We shall assume that V is stationary, i.e. it is started from $V_1 \sim \pi$. Stationarity is not really required in what follows and is solely a matter of aesthetics (e.g. $\mathbb{P}(V_1 = j) = C/j^2$ will work as well).

Let $a_0 = 0$, $a_i = 8 \sum_{j=1}^i (1/9)^j$, $i = 1, 2, \dots$ and set $A_i = [a_{i-1}, a_i)$, $i \geq 1$. Denote by $\ell_i = 8(1/9)^i$ the length of the interval A_i and note that $[0, 1) = \cup_{i=1}^{\infty} A_i$.

Now consider the observation density

$$p((u, v), y) = \mathbf{1}_{\{y \in [0, 1]\}} \mathbf{1}_{\{u \notin \cup_{i=1}^v A_i\}} + \sum_{i=1}^v \ell_i^{-1} \mathbf{1}_{\{(u, y) \in A_i \times A_i\}}.$$

As we show below, the MAP estimates of $U_{1:n}$ and $V_{1:n}$ are given by⁴:

$$\begin{aligned} \hat{U}_m^n &= \sum_{j=1}^{\infty} a_{j-1} \mathbf{1}_{\{Y_m \in A_j\}}, \quad m = 1, \dots, n \\ \hat{V}_m^n &= \begin{cases} 2 & j^*(n) = 1 \\ j^*(n) & j^*(n) > 1 \end{cases} \end{aligned} \quad (2.4)$$

where $j^*(n) := \max \left\{ j : \sum_{k=1}^n \mathbf{1}_{\{Y_k \in A_j\}} > 0 \right\}$. Since all A_j 's have positive Lebesgue measure, $j^*(n) \nearrow \infty$ as $n \rightarrow \infty$, and consequently, for any fixed $m \geq 1$,

$$\lim_{n \rightarrow \infty} \hat{V}_m^n = \lim_{n \rightarrow \infty} j^*(n) = \infty, \quad \mathbb{P} - a.s.$$

Before proving (2.4), we shall briefly explain why the optimal path of such a form should be anticipated. Note that since U_i 's are uniformly distributed in $[0, 1]$, the choice of \hat{U}_i^n 's influences the likelihood (1.1) only through the observation densities. More precisely, whenever $\{Y_m \in A_i\}$ is observed, the maximal gain of ℓ_i^{-1} is obtained if $\hat{U}_m^n \in A_i$ and $\hat{V}_m^n \geq i$ are chosen. On the other hand, the transition probabilities of (2.2) favor pathes $\hat{V}_{1:n}^n$ without jumps. Hence the optimal path $\hat{V}_{1:n}^n$ should be constant and large enough to allow the access to the narrowest A_i visited by Y_m 's so far, i.e. greater or equal to $j^*(n)$. But if constant $\hat{V}_{1:n}^n$ is chosen, it cannot be too large, as this would decrease the likelihood through the term $\pi(\hat{V}_1^n)$, due to the fast tail decay of the initial distribution π . This heuristics is implemented by an appropriate balancing between all the ingredients of the model.

⁴ The choice of \hat{U}_m^n is not unique, unless the lexicographic order is imposed: e.g. $\hat{U}_m^n := Y_m$ yields the same value of the likelihood.

Let's first check (2.4) in the case $j^*(n) > 1$. To this end, consider the ratio

$$\frac{L_n((u_{1:n}, v_{1:n}), Y_{1:n})}{L_n((\hat{U}_{1:n}^n, \hat{V}_{1:n}^n), Y_{1:n})} = \frac{\pi(v_1)}{\pi(j^*(n))} \prod_{m=2}^n \frac{P(v_{m-1}, v_m)}{P(j^*(n), j^*(n))} \prod_{m=1}^n \frac{p((u_m, v_m), Y_m)}{p((\hat{U}_m^n, j^*(n)), Y_m)}, \quad (2.5)$$

for an arbitrary $u_{1:n}$ and $v_{1:n}$. Let N be the number of jumps in $v_{1:n}$ and $v^*(n) = \max_{k=1, \dots, n} v_k$. Note that $P(v_{m-1}, v_m) = 1 - \varepsilon$, when $v_{m-1} = v_m$ and $P(v_{m-1}, v_m) \leq \varepsilon$ otherwise. Hence, as $P(j^*(n), j^*(n)) = 1 - \varepsilon$,

$$\prod_{m=2}^n \frac{P(v_{m-1}, v_m)}{P(j^*(n), j^*(n))} \leq \left(\frac{\varepsilon}{1 - \varepsilon} \right)^N.$$

Further, note that on the event $\{Y_m \in A_i\}$, $p((u_m, v_m), Y_m) \leq 1 \vee \ell_i^{-1} = \ell_i^{-1}$ and $p((\hat{U}_m^n, j^*(n)), Y_m) = \ell_i^{-1}$ and thus

$$\frac{p((u_m, v_m), Y_m)}{p((\hat{U}_m^n, j^*(n)), Y_m)} \leq 1.$$

Moreover, on $\{Y_m \in A_{j^*(n)}\}$,

$$\frac{p((u_m, v_m), Y_m)}{p((\hat{U}_m^n, j^*(n)), Y_m)} \leq \frac{\mathbf{1}_{\{v^*(n) < j^*(n)\}} + \ell_{j^*(n)}^{-1} \mathbf{1}_{\{v^*(n) \geq j^*(n)\}}}{\ell_{j^*(n)}^{-1}} \leq \frac{\ell_{v^*(n) \wedge j^*(n)}^{-1}}{\ell_{j^*(n)}^{-1}}.$$

Plugging these inequalities into (2.5), we get:

$$\begin{aligned} \frac{L_n((u_{1:n}, v_{1:n}), Y_{1:n})}{L_n((\hat{U}_{1:n}^n, \hat{V}_{1:n}^n), Y_{1:n})} &\leq \frac{\pi(v_1)}{\pi(j^*(n))} \left(\frac{\varepsilon}{1 - \varepsilon} \right)^N \frac{\ell_{v^*(n) \wedge j^*(n)}^{-1}}{\ell_{j^*(n)}^{-1}} \\ &= \frac{\pi(v_1)}{\pi(v^*(n))} \tilde{\varepsilon}^N \frac{\pi(v^*(n))}{\pi(j^*(n))} \frac{\ell_{v^*(n) \wedge j^*(n)}^{-1}}{\ell_{j^*(n)}^{-1}}, \end{aligned} \quad (2.6)$$

where $\tilde{\varepsilon} := \varepsilon/(1 - \varepsilon)$ is set for brevity. Since $N \geq v^*(n) - v_1$,

$$\begin{aligned} \frac{\pi(v_1)}{\pi(v^*(n))} \tilde{\varepsilon}^N &\leq \frac{\pi(v_1)}{\pi(v^*(n))} \tilde{\varepsilon}^{v^*(n) - v_1} \\ &\leq \left(\frac{v^*(n)}{v_1} \right)^2 \frac{1 + \left(\frac{v_1}{v_1 + 1} \right)^2}{1 + \left(\frac{v^*(n)}{v^*(n) + 1} \right)^2} \tilde{\varepsilon}^{v^*(n) - v_1} \leq \left(\frac{v^*(n)}{v_1} \right)^2 \tilde{\varepsilon}^{v^*(n) - v_1}, \end{aligned}$$

where in the second inequality we used the expression for $\pi(j)$, $j > 1$ from (2.3). In fact, the inequality is true also for $v^*(n) = v_1 = 1$, as both the right and the left hand side turn to 1, and for $v^*(n) > v_1 = 1$, as $\pi(1)$ is less than $C \frac{1}{j^2} \left(1 + \left(\frac{j}{j+1} \right)^2 \right)$, evaluated at $j := 1$.

The function $x \mapsto x^2 \tilde{\varepsilon}^x$ attains its maximum at $x^* = 2/\log \tilde{\varepsilon}^{-1}$ and is strictly decreasing on (x^*, ∞) . Hence with $\tilde{\varepsilon} < e^{-2}$, that is with $\varepsilon < e^{-2}/(1 + e^{-2})$, for any $y > x \geq 1$, $(y/x)^2 \tilde{\varepsilon}^{y-x} < 1$ and hence

$$\frac{\pi(v_1)}{\pi(v^*(n))} \tilde{\varepsilon}^N \leq 1. \quad (2.7)$$

The equality holds if and only if $v_{1:n}$ is a constant path, i.e. $v_m = v^*(n)$ for all $m = 1, \dots, n$.

Further, if $v^*(n) \leq j^*(n)$,

$$\begin{aligned} \frac{\pi(v^*(n))}{\pi(j^*(n))} \frac{\ell_{v^*(n) \wedge j^*(n)}^{-1}}{\ell_{j^*(n)}^{-1}} &\leq \left(\frac{j^*(n)}{v^*(n)} \right)^2 \frac{1 + \left(\frac{v^*(n)}{v^*(n)+1} \right)^2}{1 + \left(\frac{j^*(n)}{j^*(n)+1} \right)^2} (1/9)^{j^*(n)-v^*(n)} \\ &\leq \left(\frac{j^*(n)}{v^*(n)} \right)^2 (1/9)^{j^*(n)-v^*(n)} \leq 1, \end{aligned} \quad (2.8)$$

where the latter inequality holds since $1/9 < e^{-2}/(1 + e^{-2})$.

The sequence $\pi(j)$ attains its unique maximum at $j := 2$ and is strictly decreasing for $j \geq 2$. Hence, if $v^*(n) > j^*(n) \geq 2$,

$$\frac{\pi(v^*(n))}{\pi(j^*(n))} \frac{\ell_{v^*(n) \wedge j^*(n)}^{-1}}{\ell_{j^*(n)}^{-1}} < 1.$$

Plugging (2.7) and (2.8) into (2.6), yields the inequality for any $u_{1:n}$ and $v_{1:n}$

$$L_n((u_{1:n}, v_{1:n}), Y_{1:n}) \leq L_n((\hat{U}_{1:n}^n, \hat{V}_{1:n}^n), Y_{1:n}),$$

which saturates if and only if $v_m = j^*(n)$, $m = 1, \dots, n$, thus verifying optimality of (2.4) on the event $\{j^*(n) > 1\}$.

We shall omit the details in the case $\{j^*(n) = 1\}$, which is treated similarly: the optimal value $\hat{V}_m^n = 2$ is obtained, since $\pi(j)$ is maximal at $j = 2$. Of course, as $j^*(n)$ eventually leaves the state 1, the exact value is irrelevant for the main point of the present example, that is the divergence $\lim_{n \rightarrow \infty} \hat{V}_m^n = \infty$. ■

3. Convergence in the case of log-concave densities

In this section we establish the existence of the limit (1.2), deducing it from certain strong log-concavity properties of the densities q and p . Hereafter the following assumptions are in force:

- a1. the initial state density μ is a $C^2(\mathbb{R})$ log-concave function on \mathbb{R} and $-\log \mu(u) \geq 0$;
- a2. the hidden state transition density q is a $C^2(\mathbb{R}^2)$ log-concave function, namely ⁵
 $q(u, v) \propto e^{-\alpha(u, v)}$, where $\alpha(u, v)$ is a nonnegative twice continuously differentiable convex function on \mathbb{R}^2 ;

⁵ $f \propto g$ means that f/g is constant

- a3. the observation density p is $C^2(\mathbb{R})$ log-concave function in the first argument: $p(x, y) \propto e^{-\gamma(x, y)}$, where for each $y \in \mathbb{R}$ the function $x \mapsto \gamma(x, y)$ is nonnegative, twice continuously differentiable and strongly convex on \mathbb{R} with $x_*(y) := \operatorname{argmin}_{x \in \mathbb{R}} \gamma(x, y) \in (-\infty, \infty)$ and

$$\frac{\partial^2}{\partial x^2} \gamma(x, y) \geq \kappa > 0, \quad \forall x, y \in \mathbb{R},$$

with a constant κ .

- a4. for some constant C ,

$$-\overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log L_n(X_{1:n}, Y_{1:n}) \leq C, \quad \mathbb{P} - a.s.$$

- a5. there is a non-decreasing function $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$, growing to $+\infty$ not faster than a polynomial, such that for all $M > 0$

$$\alpha(x, y) \leq M \implies \left| \frac{\partial^2}{\partial x \partial y} \alpha(x, y) \right| \leq g(M), \quad \forall x, y \in \mathbb{R}.$$

Remark 3.1. The log-concavity assumptions (a1)-(a3) are quite restrictive. For example, if $Y_n = h(X_n) + w_n$ with $w_n \sim N(0, 1)$, then

$$\frac{\partial^2}{\partial x^2} \gamma(x, y) = \frac{1}{2} \frac{\partial^2}{\partial x^2} (y - h(x))^2 = (h'(x))^2 - (y - h(x))h''(x),$$

which typically will not admit the uniform lower bound of (a3), unless h is linear, i.e. $h''(x) \equiv 0$.

If the assumption (a3) is satisfied, it implies $\gamma_*(y) := \gamma(x_*(y), y) \in (-\infty, \infty)$, for all $y \in \mathbb{R}$ and, moreover,

$$\gamma(x, y) - \gamma_*(y) \geq \frac{1}{2} \kappa (x - x_*)^2, \quad \forall x, y \in \mathbb{R}, \quad (3.1)$$

which is essential to our approach.

Assuming that $-\log \mu(u)$, $\alpha(u, v)$ and $\gamma(x, y)$ are nonnegative is equivalent to assuming that they are lower bounded by a constant, i.e. that the corresponding densities are bounded.

The assumption (a4) is typically satisfied, if the state process X is positively recurrent (explicit recurrence tests can be found in [Meyn and Tweedie \(2009\)](#); see also [Genon-Catalot, Jeantheau and Larédo \(2000\)](#)). Finally, (a5) is a technical assumption, which is satisfied in most models of practical interest.

Example 3.1. All the above assumptions are satisfied for the linear HMM

$$\begin{aligned} X_n &= aX_{n-1} + v_n, \quad n \geq 1 \\ Y_n &= bX_n + w_n, \end{aligned}$$

where $|a| < 1$ and $b \neq 0$ are constants and $v = (v_n)_{n \geq 1}$ and $w = (w_n)_{n \geq 1}$ are independent sequences of i.i.d random variables with

$$X_0, v_n \sim f_v(x) \propto e^{-|x|^{2+\delta}} \quad \text{and} \quad w_n \sim f_w(x) \propto e^{-x^2(1+c|x|^{\delta'})}$$

for some $\delta \geq 0$ and $\delta' \geq 0$, $c \geq 0$. ■

Theorem 3.1. *The limit in (1.2) exists \mathbb{P} -a.s.*

Proof. To keep the notations simple, we shall prove the convergence in (1.2) for $m = 1$, i.e. the limit $\lim_{n \rightarrow \infty} \hat{X}_1^n$ exists \mathbb{P} -a.s. As will be clear from the proof below, the same arguments imply convergence $\lim_{n \rightarrow \infty} \hat{X}_i^n$ for any $i \leq m$ and hence of (1.2) for any fixed $m \geq 1$.

To check $\lim_{n \rightarrow \infty} \hat{X}_1^n$, \mathbb{P} -a.s., we shall show that on a set of probability one the series

$$\hat{X}_1^n = \hat{X}_1^1 + \sum_{i=2}^n (\hat{X}_1^i - \hat{X}_1^{i-1})$$

is convergent. The proof hinges on the system of inequalities (3.6) and (3.7), which stem from the log-concavity properties assumed in (a1)-(a3). A pigeonhole principle type of argument (Lemma A.1) shows that a sequence satisfying such inequalities must decay at least polynomially backward in time, which in turn yields the desired conclusion.

To this end, introduce⁶

$$\begin{aligned} h_n(x_{1:n}) &:= -\log L_n(x_{1:n}, Y_{1:n}) \\ &= -\log \mu(x_1) + \gamma(x_1, Y_1) + \sum_{m=2}^n (\alpha(x_{m-1}, x_m) + \gamma(x_m, Y_m)). \end{aligned} \quad (3.2)$$

By assumptions (a1)-(a3), $\lim_{R \rightarrow \infty} \inf_{\|x_{1:n}\|=R} h_n(x_{1:n}) \rightarrow \infty$, and for any $n \geq 1$ the function

$$x_{1:n} \mapsto h_n(x_{1:n}) + \alpha(x_n, u) \quad (3.3)$$

attains its global minimum at

$$\tilde{X}_{1:n}^n(u) := \operatorname{argmin}_{x_{1:n}} (h_n(x_{1:n}) + \alpha(x_n, u)), \quad u \in \mathbb{R}.$$

The Hessian matrix of the function defined in (3.3) is positive definite uniformly over $x_{1:n} \in \mathbb{R}^n$ and hence the minimum is unique and $\tilde{X}_{1:n}^n(u)$ is the solution of

$$\operatorname{grad}(h_n(x_{1:n}) + \alpha(x_n, u)) = 0.$$

The Jacobian matrix of the function in the left hand side of this equation with respect to the vector $x_{1:n}$ coincides with the aforementioned Hessian matrix and hence is invertible

⁶for $k > \ell$, $\sum_{i=k}^{\ell} \dots = 0$ is understood

at any $u \in \mathbb{R}$. Thus by the implicit function theorem $u \mapsto \tilde{X}_{1:n}^n(u)$ is continuously differentiable on \mathbb{R} .

The usual dynamical programming argument yields the following chain rules:

$$\begin{aligned}\tilde{X}_j^n(x) &= \tilde{X}_j^m(\tilde{X}_{m+1}^n(x)), \quad x \in \mathbb{R}, \quad j < n, \quad m = j, \dots, n, \\ \hat{X}_j^n &= \tilde{X}_j^m(\hat{X}_{m+1}^n).\end{aligned}\tag{3.4}$$

Hence for $j < n$, and $j \leq m < n$,

$$\begin{aligned}\hat{X}_j^{n+1} - \hat{X}_j^n &= \tilde{X}_j^m(\hat{X}_{m+1}^{n+1}) - \tilde{X}_j^m(\hat{X}_{m+1}^n) \\ &= (\hat{X}_{m+1}^{n+1} - \hat{X}_{m+1}^n) \int_0^1 \frac{\partial}{\partial s} \tilde{X}_j^m(s\hat{X}_{m+1}^{n+1} + (1-s)\hat{X}_{m+1}^n) ds.\end{aligned}\tag{3.5}$$

The following lemma is the key to a bound on the integrand in (3.5):

Lemma 3.1. Assume (a1)-(a3), then for $j = 1, \dots, n-1$,

$$\left\| \frac{\partial}{\partial x} \tilde{X}_{1:j}^n(x) \right\|^2 \leq \frac{2}{\kappa} \left| \mathcal{D}_{12} \alpha(\tilde{X}_j^n(x), \tilde{X}_{j+1}^n(x)) \frac{\partial}{\partial x} \tilde{X}_{j+1}^n(x) \frac{\partial}{\partial x} \tilde{X}_j^n(x) \right|,\tag{3.6}$$

and

$$\left\| \frac{\partial}{\partial x} \tilde{X}_{1:n}^n(x) \right\|^2 \leq \frac{2}{\kappa} \left| \mathcal{D}_{12} \alpha(\tilde{X}_n^n(x), x) \frac{\partial}{\partial x} \tilde{X}_n^n(x) \right|,\tag{3.7}$$

where $\mathcal{D}_{12} \alpha(x, y) := \frac{\partial^2}{\partial x \partial y} \alpha(x, y)$, and κ is as in Assumption (a3).

Proof. Recall that the function (3.3) is convex with the Hessian, greater than κ times identity matrix, with respect to the positive definite ordering. Hence, for any $1 \leq j < n$ and $u, v \in \mathbb{R}$, by (3.1),

$$\frac{\kappa}{2} \left\| \tilde{X}_{1:j}^j(v) - \tilde{X}_{1:j}^j(u) \right\|^2 \leq h_j(\tilde{X}_{1:j}^j(v)) + \alpha(\tilde{X}_{1:j}^j(v), u) - h_j(\tilde{X}_{1:j}^j(u)) - \alpha(\tilde{X}_{1:j}^j(u), u),$$

since, by definition, the minimum of $h_j(x_{1:j}) + \alpha(x_j, u)$ over $x_{1:j}$ is attained at $\tilde{X}_{1:j}^j(u)$. Further, by definition of $\tilde{X}_{1:j}^j(v)$,

$$h_j(\tilde{X}_{1:j}^j(v)) + \alpha(\tilde{X}_{1:j}^j(v), v) \leq h_j(\tilde{X}_{1:j}^j(u)) + \alpha(\tilde{X}_{1:j}^j(u), v),$$

which gives

$$\frac{\kappa}{2} \left\| \tilde{X}_{1:j}^j(v) - \tilde{X}_{1:j}^j(u) \right\|^2 \leq -\alpha(\tilde{X}_{1:j}^j(v), v) + \alpha(\tilde{X}_{1:j}^j(u), v) + \alpha(\tilde{X}_{1:j}^j(v), u) - \alpha(\tilde{X}_{1:j}^j(u), u).\tag{3.8}$$

Plugging $v := \tilde{X}_{j+1}^n(x+h)$ and $u := \tilde{X}_{j+1}^n(x)$ with $x \in \mathbb{R}$ and using the chain rule (3.4), we get

$$\begin{aligned} \frac{\kappa}{2} \|\tilde{X}_{1:j}^n(x+h) - \tilde{X}_{1:j}^n(x)\|^2 &\leq -\alpha(\tilde{X}_j^n(x+h), \tilde{X}_{j+1}^n(x+h)) + \\ &\quad \alpha(\tilde{X}_j^n(x), \tilde{X}_{j+1}^n(x+h)) + \alpha(\tilde{X}_j^n(x+h), \tilde{X}_{j+1}^n(x)) - \alpha(\tilde{X}_j^n(x), \tilde{X}_{j+1}^n(x)). \end{aligned}$$

Since all the functions appearing in the latter inequality are twice continuously differentiable, dividing by h^2 and taking $h \rightarrow 0$ gives the bound (3.6). Similarly, with $j := n$, $v := x+h$ and $u := x$, (3.8) yields (3.7). \square

By assumption (a4),

$$\Omega' := \left\{ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=2}^n \left(\alpha(X_{j-1}, X_j) + \gamma(X_j, Y_j) \right) \leq C \right\},$$

is an event of full probability and hence it is enough to verify the claimed convergence for all $\omega \in \Omega'$. Clearly, for an $\omega \in \Omega'$,

$$-\log \mu(X_1) + \gamma(X_1, Y_1) + \sum_{j=2}^n \left(\alpha(X_{j-1}, X_j) + \gamma(X_j, Y_j) \right) \leq 2Cn, \quad \forall n \geq N(\omega),$$

for an integer $N(\omega) < \infty$. Then $\hat{X}_{1:n}^n$, being a minimizer, a fortiori satisfies:

$$-\log \mu(\hat{X}_1^n) + \gamma(\hat{X}_1^n, Y_1) + \sum_{j=2}^n \left(\alpha(\hat{X}_{j-1}^n, \hat{X}_j^n) + \gamma(\hat{X}_j^n, Y_j) \right) \leq 2Cn, \quad \forall n \geq N. \quad (3.9)$$

Hence for a large fixed constant $M > 4C$ and any $n \geq N$,

$$\#\{j : \alpha(\hat{X}_{j-1}^n, \hat{X}_j^n) + \gamma(\hat{X}_j^n, Y_j) > M\} \leq \frac{2Cn}{M} =: \rho n.$$

Similarly,

$$\#\{j : \alpha(\hat{X}_{j-1}^{n+1}, \hat{X}_j^{n+1}) + \gamma(\hat{X}_j^{n+1}, Y_j) > M\} \leq \frac{2C(n+1)}{M} = \rho(n+1).$$

Then there is an index $m \in [n - 2\rho n, n]$, such that

$$\alpha(\hat{X}_{m-1}^n, \hat{X}_m^n) + \gamma(\hat{X}_m^n, Y_m) \leq M, \quad \text{and} \quad \alpha(\hat{X}_{m-1}^{n+1}, \hat{X}_m^{n+1}) + \gamma(\hat{X}_m^{n+1}, Y_m) \leq M,$$

and, by the assumption (a3),

$$\begin{aligned} |\hat{X}_m^{n+1} - \hat{X}_m^n| &\leq |\hat{X}_m^{n+1} - \operatorname{argmin}_{x \in \mathbb{R}} \gamma(x, Y_m)| + |\hat{X}_m^n - \operatorname{argmin}_{x \in \mathbb{R}} \gamma(x, Y_m)| \\ &\leq \sqrt{\frac{2}{\kappa}} \left(\gamma(\hat{X}_m^{n+1}, Y_m) - \gamma_*(Y_m) \right)^{1/2} + \sqrt{\frac{2}{\kappa}} \left(\gamma(\hat{X}_m^n, Y_m) - \gamma_*(Y_m) \right)^{1/2} \\ &\leq \sqrt{\frac{2}{\kappa} \gamma(\hat{X}_m^n, Y_m)} + \sqrt{\frac{2}{\kappa} \gamma(\hat{X}_m^{n+1}, Y_m)} \leq \sqrt{\frac{8M}{\kappa}}. \end{aligned}$$

Plugging this estimate into (3.5), we get (for $j := 1$)

$$|\hat{X}_1^{n+1} - \hat{X}_1^n| \leq \sqrt{\frac{8M}{\kappa}} \int_0^1 \left| \frac{\partial}{\partial s} \tilde{X}_1^{m-1} \left(s\hat{X}_m^{n+1} + (1-s)\hat{X}_m^n \right) \right| ds. \quad (3.10)$$

Introduce

$$\begin{aligned} \check{X}_m^m(s) &:= s\hat{X}_m^{n+1} + (1-s)\hat{X}_m^n, \\ \check{X}_j^m(s) &:= \tilde{X}_j^{m-1}(\check{X}_m^m(s)), \quad j = 1, \dots, m-1, \end{aligned}$$

and define

$$\begin{aligned} c_j(s) &:= \frac{2}{\kappa} \left| \mathcal{D}_{12} \alpha(\check{X}_j^m(s), \check{X}_{j+1}^m(s)) \right|, \quad j < m, \\ b_j(s) &:= \left| \frac{\partial}{\partial x} \tilde{X}_j^{m-1}(\check{X}_m^m(x)) \right|_{x := s}. \end{aligned}$$

Then from (3.6) and (3.7) (the dependence on s is now omitted for brevity)

$$\sum_{i=1}^j b_i^2 \leq c_j b_j b_{j+1}, \quad j < m-1 \quad (3.11)$$

$$\sum_{i=1}^{m-1} b_i^2 \leq c_{m-1} b_{m-1}, \quad (3.12)$$

and (3.10) reads:

$$|\hat{X}_1^{n+1} - \hat{X}_1^n| \leq \sqrt{\frac{8M}{\kappa}} \int_0^1 b_1(s) ds. \quad (3.13)$$

Lemma 3.2. For any $s \in [0, 1]$, $x > 0$, and $g(\cdot)$ as in (a5),

$$\#\{j < m : c_j(s) > \frac{2}{\kappa} g(x)\} \leq \frac{4C}{x(1-2\rho)} m. \quad (3.14)$$

Proof. The function $u \mapsto \min_{x_{1:n}} \left(h_n(x_{1:n}) + \alpha(x_n, u) \right)$ is convex and hence

$$\begin{aligned}
\sum_{j=2}^m \alpha(\check{X}_{j-1}^m, \check{X}_j^m) &\leq h_{m-1}(\check{X}_{1:m-1}^m) + \alpha(\check{X}_{m-1}^m, \check{X}_m^m) \\
&= \min_{x_{1:m-1}} \left(h_{m-1}(x_{1:m-1}) + \alpha(x_{m-1}, s\hat{X}_m^{n+1} + (1-s)\hat{X}_m^n) \right) \\
&\leq s \min_{x_{1:m-1}} \left(h_{m-1}(x_{1:m-1}) + \alpha(x_{m-1}, \hat{X}_m^{n+1}) \right) \\
&\quad + (1-s) \min_{x_{1:m-1}} \left(h_{m-1}(x_{1:m-1}) + \alpha(x_{m-1}, \hat{X}_m^n) \right) \\
&= s \left(h_{m-1}(\hat{X}_{1:m-1}^{n+1}) + \alpha(\hat{X}_{m-1}^{n+1}, \hat{X}_m^{n+1}) \right) \\
&\quad + (1-s) \left(h_{m-1}(\hat{X}_{1:m-1}^n) + \alpha(\hat{X}_{m-1}^n, \hat{X}_m^n) \right) \\
&\leq 2C(n+1),
\end{aligned}$$

where the latter inequality follows from (3.9). Hence

$$\#\{j \leq m : \alpha(\check{X}_{j-1}^m, \check{X}_j^m) > x\} \leq \frac{2C(n+1)}{x},$$

and, since $m \geq (1-2\rho)n$, (3.14) follows from the assumption (a5). \square

Now by Corollary A.1 in the Appendix, applied to (3.11)-(3.12) and (3.14), for any $\beta > 1$, there is a constant C_β , such that

$$b_1 \leq C_\beta m^{-\beta} \leq C_\beta (1-2\rho)^{-\beta} n^{-\beta}, \quad (3.15)$$

for all sufficiently large n , and thus, by (3.13), the sequence $|\hat{X}_1^{n+1} - \hat{X}_1^n|$, $n \geq 1$ is summable, which verifies the existence of the limit (1.2). \square

Remark 3.2. When the hidden state process is a Gaussian autoregression, i.e. when $\alpha(x, y) = \frac{1}{2}(y - bx)^2$ with a constant $b \neq 0$, $|\mathcal{D}_{12}\alpha(x, y)| \equiv b$ and Lemma A.1 (1) implies exponential bound in (3.15), confirming the results deducible from the Kalman linear filtering theory.

4. Concluding remarks

As indicated by the examples of Section 2 and the partial results of Theorem 3.1, the convergence in (1.2) appears to be a non-trivial issue. Analogous problems have been discussed in the engineering literature. In fact, the MAP path estimation can be viewed as an optimal control problem, in which one is required to minimize the cost functional $h_n(x_{1:n})$ defined in (3.2), where the term $\alpha(x_{m-1}, x_m)$ is interpreted as the cost incurred

by the control effort (needed to move from x_{m-1} to x_m) and $\gamma(x_m, Y_m)$ is the cost paid for the deviation of the state from Y_m . This setting appears in R. Bellman's book [Bellman \(1957\)](#) Ch. I, Sec. 1.7. as the “smoothing” problem and in the control literature is often referred to as the *tracking* problem. From the control theory perspective, the existence of the limit in (1.2) means that the optimal control and the corresponding optimal trajectory cease to depend on the future values of the exogenous signal Y .

Among other related questions, the convergence (1.2) of the optimal trajectory is a part of the “asymptotic control theory” program, initiated by R. Kalman, R. Bellman and R. Bucy yet on the dawn of the modern control theory. In the linear state/quadratic cost (LQ) setting of R. Kalman, the control problem admits an elegant closed form solution for each fixed horizon n and the study of the limit (1.2) reduces to the stability analysis of the associated Riccati equation (a comprehensive treatment of the LQ problem can be found in e.g. [Kwakernaak and Sivan \(1972\)](#)).

To the best of our knowledge, asymptotic analysis beyond the LQ case has been carried out only for a limited number of nonlinear models. [Bellman and Bucy \(1964\)](#) found a remarkable explicit solution to a quite general scalar continuous-time control problem, amenable to asymptotic analysis. A vector control problem with linear state dynamics and convex costs was studied in [Bucy \(1966\)](#).

While much progress has been achieved in the optimal control theory on the *infinite horizon* (see e.g. [Carlson, Haurie and Leizarowitz \(1991\)](#), [Zaslavski \(2006\)](#)), we were not able to track any results, directly applicable to the question under consideration.

Another possible connection, remaining elusive at the moment, is to the stability theory of nonlinear filtering equation, developed during the last decade (see e.g. the survey [Chigansky, Liptser and Van Handel \(2009\)](#)).

Appendix A: A supporting lemma

Lemma A.1. *Consider the system of inequalities:*

$$\begin{aligned} \sum_{i=1}^j b_i^2 &\leq b_j b_{j+1} c_j, \quad j = 1, \dots, n-1 \\ \sum_{i=1}^n b_i^2 &\leq b_n c_n \end{aligned} \tag{A.1}$$

where b_i and c_i , $i = 1, \dots, n$ are nonnegative real numbers and let θ and θ' be arbitrary positive constants.

1. If $c_i \leq \theta$, $i = 1, \dots, n$ then

$$b_1 \leq \sqrt{\theta e} \exp\left(-\frac{n}{2e(\theta^2 \vee \theta)}\right), \quad \text{for } n \geq \theta^2 e. \tag{A.2}$$

2. If for a non-decreasing nonnegative function $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$,

$$\#\{i \leq n : c_i \geq g(x)\} \leq \frac{\theta n}{x}, \quad \forall x > 0, \quad (\text{A.3})$$

and $c_n \leq \theta'$, then for any $p \in (0, 1)$ and $\ell > \theta$,

$$b_1 \leq \sqrt{g(\ell)n^{-p\ell/(4\theta)}} \quad \text{for } n > \left(\frac{\ell(\theta'^2 \vee g(\ell))}{\theta} \right)^{1/(1-p)}. \quad (\text{A.4})$$

3. If only (A.3) holds, then for any $p \in (0, 1)$ and $\ell > \theta$,

$$b_1 \leq g(2\theta n) \sqrt{g(\ell)n^{-p\ell/(4\theta)}} \quad \text{for } n > \left(\frac{\ell(1 \vee g(\ell))}{\theta} \right)^{1/(1-p)}. \quad (\text{A.5})$$

Proof.

1) The second inequality in (A.1) and $c_n \leq \theta$, imply $b_n^2 \leq b_n \theta$ and in turn $b_1^2 + \dots + b_n^2 \leq \theta^2$. Fix a constant $\eta \in (0, 1)$ and let $m_1 := \lfloor \theta^2/\eta \rfloor$. Then at most half of b_i 's with $i \in [n - 2m_1, n]$ are greater than $\sqrt{\eta}$ and hence there is an index $k_1 \in [n - 2m_1, n]$, such that $b_{k_1} \leq \sqrt{\eta}$ and $b_{k_1+1} \leq \sqrt{\eta}$. The inequality corresponding to $j := k_1$ in (A.1) then gives the bound $b_1^2 + \dots + b_{k_1}^2 \leq b_{k_1} b_{k_1+1} c_{k_1} \leq \eta \theta$.

Similarly, let $m_2 := \lfloor \theta/\eta \rfloor$, then there is an index $k_2 \in [k_1 - 2m_2 : k_1]$, such that $b_{k_2} \leq \eta$ and $b_{k_2+1} \leq \eta$ and, again, applying (A.1), $b_1^2 + \dots + b_{k_2}^2 \leq b_{k_2} b_{k_2+1} c_{k_2} \leq \eta^2 \theta$. This argument can be iterated at least

$$\left\lfloor \frac{n}{2(m_1 \vee m_2)} \right\rfloor = \left\lfloor \frac{\eta n}{2(\theta^2 \vee \theta)} \right\rfloor$$

times and thus

$$b_1^2 \leq \theta \eta^{\lfloor \frac{1}{2} \eta n / (\theta^2 \vee \theta) \rfloor} \leq \frac{\theta}{\eta} \left((\eta^\eta)^{\frac{1}{2}/(\theta^2 \vee \theta)} \right)^n.$$

The best rate is obtained at $\eta := e^{-1}$, which yields the bound (A.2).

2) For a fixed $\ell \geq \theta$, by (A.3)

$$\#\{i \leq n : c_i \geq g(\ell)\} \leq \frac{\theta n}{\ell} := rn, \quad (\text{A.6})$$

and thus at least half of c_i 's with $i \in [n - 2rn, n]$ do not exceed $g(\ell)$. Fix a constant $p \in (0, 1)$ and let $\eta := n^{-p/2}$. Suppose that for all $i \in [n - 2rn, n]$, such that $c_i \leq g(\ell)$, either $b_i \geq \eta$ or $b_{i+1} \geq \eta$ or both; then

$$\#\{i \in [n - 2rn : n] : b_i \geq \eta\} \geq rn.$$

But on the other hand, by the second inequality in (A.1) and as $c_n \leq \theta'$, $b_n^2 \leq b_n \theta'$ and $b_1^2 + \dots + b_n^2 \leq \theta'^2$, and hence

$$\#\{i \in [n - 2rn : n] : b_i \geq \eta\} \leq \frac{\theta'^2}{\eta^2} = \theta'^2 n^p.$$

This contradicts the previous estimate if n is large enough, namely, if $n > (\ell\theta'^2/\theta)^{1/(1-p)}$. Thus for such n , there is an index $m_1 \in [n - 2rn : n]$, such that $b_{m_1} \leq \eta$, $b_{m_1+1} \leq \eta$ and $c_{m_1} \leq g(\ell)$.

Now by the inequality in (A.1), corresponding to $j := m_1$,

$$b_1^2 + \dots + b_{m_1}^2 \leq b_{m_1} b_{m_1+1} c_{m_1} \leq \eta^2 g(\ell), \quad (\text{A.7})$$

for which the above consideration can be repeated. Namely, by (A.6), there are at least rn indices $i \in [m_1 - 2rn, m_1]$, for which $c_i \leq g(\ell)$. Suppose that for all of them either $b_i \geq \eta^2$ or $b_{i+1} \geq \eta^2$ or both, then

$$\#\{i \in [m_1 - 2rn : m_1] : b_i \geq \eta^2\} \geq rn,$$

while (A.7) implies

$$\#\{i \in [m_1 - 2rn : m_1] : b_i \geq \eta^2\} \leq \frac{\eta^2 g(\ell)}{\eta^4} = n^p g(\ell),$$

which is a contradiction for n large enough, i.e. for $n > (\ell g(\ell)/\theta)^{1/(1-p)}$. Hence there is an $m_2 \in [m_1 - 2rn : m_1]$, such that $b_{m_2} \leq \eta^2$, $b_{m_2+1} \leq \eta^2$ and $c_{m_2} \leq g(\ell)$ and thus by (A.1)

$$b_1^2 + \dots + b_{m_2}^2 \leq \eta^4 g(\ell).$$

This argument can be iterated for at least $\lfloor 1/(2r) \rfloor$ times, which yields the bound:

$$b_1^2 \leq g(\ell) (\eta^{1/2r})^2 = g(\ell) n^{-p\ell/2\theta}.$$

3) Note that $b'_i := b_i/g(2\theta n)$, $i = 1, \dots, n$ satisfy the inequalities (A.1) with c_i 's replaced with $c'_i := c_i$, $i = 1, \dots, n-1$ and $c'_n := c_n/g(2\theta n)$. By (A.3),

$$\#\{i \leq n : c_i \geq g(2\theta n)\} \leq \frac{\theta n}{2\theta n} = 1/2,$$

i.e. all c_i 's are less than $g(2\theta n)$ and, in particular, $c_n \leq g(2\theta n)$, that is $c'_n \leq 1$. Moreover, assuming that $g(2\theta n) \geq 1$,

$$\#\{i \leq n : c'_i \geq g(x)\} \leq \#\{i \leq n : c_i \geq g(x)\} \leq \frac{\theta n}{x}, \quad \forall x > 0.$$

Hence by (A.4)

$$b'_1 \leq \sqrt{g(\ell)} n^{-p\ell/(2\theta)} \quad \text{for } n > \left(\frac{\ell(1 \vee g(\ell))}{\theta} \right)^{1/(1-p)},$$

which in turn gives (A.5). □

Corollary A.1. *Under the assumption (A.3) with $g(\cdot)$ growing to $+\infty$ not faster than a polynomial, for any $\beta > 1$, there is a constant C_β , such that*

$$b_1 \leq C_\beta n^{-\beta},$$

for all sufficiently large n .

Proof. Follows from (3) of Lemma A.1. □

Acknowledgements

The authors are grateful to M. Margaliot and R. Van Handel for the enlightening discussion on the issues raised in this paper and appreciate the thorough proofreading by the referees.

P.Chigansky was supported by ISF grant 314/09; Yaacov Ritov was supported by ISF grant

References

- BELLMAN, R. (1957). *Dynamic programming*. Princeton University Press, Princeton, N. J.
- BELLMAN, R. and BUCY, R. (1964). Asymptotic control theory. *J. Soc. Indust. Appl. Math. Ser. A Control* **2** 11–18.
- BUCY, R. S. (1966). New results in asymptotic control theory. *SIAM J. Control* **4** 397–402.
- CALIEBE, A. (2006). Properties of the maximum a posteriori path estimator in hidden Markov models. *IEEE Trans. Inform. Theory* **52** 41–51.
- CALIEBE, A. and RÖSLER, U. (2002). Convergence of the maximum a posteriori path estimator in hidden Markov models. *IEEE Trans. Inform. Theory* **48** 1750–1758.
- CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in hidden Markov models*. *Springer Series in Statistics*. Springer, New York.
- CARLSON, D., HAURIE, A. and LEIZAROWITZ, A. (1991). *Infinite horizon optimal control: deterministic and stochastic systems*, 2nd ed. Springer-Verlag.
- CHIGANSKY, P., LIPTSER, R. and VAN HANDEL, R. (2009). Intrinsic methods in filter stability. In *Handbook of Nonlinear Filtering* (D. Crisan and B. Rozovskii, eds.) Oxford University Press to appear.
- GENON-CATALOT, V., JEANTHEAU, T. and LARÉDO, C. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli* **6** 1051–1079.
- KOGAN, J. A. (1996). Hidden Markov models estimation via the most informative stopping times for the Viterbi algorithm. In *Image models (and their speech model cousins)* (Minneapolis, MN, 1993/1994). *IMA Vol. Math. Appl.* **80** 115–130. Springer, New York.

- KWAKERNAAK, H. and SIVAN, R. (1972). *Linear optimal control systems*. Wiley-Interscience [John Wiley & Sons], New York.
- LEMBER, J. and KOLOYDENKO, A. (2008). The adjusted Viterbi training for hidden Markov models. *Bernoulli* **14** 180–206.
- LEMBER, J. and KOLOYDENKO, A. (2010). A constructive proof of the existence of Viterbi processes. *IEEE Trans. on Information Theory* **56** 2017 - 2033.
- MEYN, S. and TWEEDIE, R. L. (2009). *Markov chains and stochastic stability*, Second ed. Cambridge University Press, Cambridge. With a prologue by Peter W. Glynn.
- ZASLAVSKI, A. J. (2006). *Turnpike properties in the calculus of variations and optimal control. Nonconvex Optimization and its Applications* **80**. Springer, New York.
- ZEITOUNI, O. and DEMBO, A. (1987). A maximum a posteriori estimator for trajectories of diffusion processes. *Stochastics* **20** 221–246.
- ZEITOUNI, O. and DEMBO, A. (1988). An existence theorem and some properties of maximum a posteriori estimators of trajectories of diffusions. *Stochastics* **23** 197–218.